# Imaging, Cataloging & Publishing the World's Vital Records

***Paul D. Abbott; FamilySearch; Salt Lake City, Utah/ USA***

## Abstract

*FamilySearch is all about preserving and publishing the world's vital records. The purpose of this paper is to describe the processes, tools and standards used by FamilySearch to capture, catalog and publish to the web hundreds of millions of images. There are five steps used by FamilySearch in publishing images; 1) identifying and gathering collection data, 2) creating and finalizing projects, 3) field capture, 4) audit and evaluation, 5) publishing, preserving and donating collections.*

*dCam-X is a tool created by FamilySearch that takes complex digital concepts and simplifies them down so that non-technical people can create publishable quality records. Managing projects, evaluating metadata, calibrating the system, taking pictures, evaluating content and shipping finished product for processing and preservation are steps greatly simplified by dCam-X. FamilySearch has implemented processes, tools and standards to ensure that the world's collections of vital records are published as accurately and as rapidly as possible.*

## Introduction

FamilySearch is all about preserving and publishing the world's vital records. These collections are gathered in from the great archives of the world along with being gathered from remote villages on the plains in Africa, from the mountains of Latin America and from the islands of the South Pacific. FamilySearch works with records that are as diverse as the cultures, locations and people of this plant. And yet, as diverse as people and cultures are, there are common events and relationships that define the human experience. There are also innate desires to preserve the history of these common events. The collections of these historical events are the vital records FamilySearch publishing.

The purpose of this paper is to provide an overview the processes, tools and standards used by FamilySearch to capture, catalog and publish to the web hundreds of millions of images. The reader will see that the key to managing high volumes of data, is having well defined standards and tools that support capturing and processing images and metadata.

## The Publishing Process

Collections processed by FamilySearch will go through a five step process starting with discovery and ending with publishing. This paper will briefly review each of the five steps of the publishing process and then go into greater detail about the tools and processes used in the field at the time the images are originally captured.

Publishing collections begins long before a camera operator shows up at an archive ready to work. There are five steps that we use in the publishing process. They are as follows:

1. Indentify and gather data about collections
2. Create and finalize a project with record custodians
3. Capture the collection
4. Audit and evaluate the images and associated metadata
5. Publish, preserve and donate the collection

*Step 1: Identify and gather information about collections and the custodians who are responsible for those collections.* The objective of this first step is to gather in and create a global view of the world's vital records. FamilySearch tries to know what records are available, the condition of the records, the types of records, where they are located and what information is available on the different records. This information allows FamilySearch to prioritize where to focus their attention. To help gather this data, FamilySearch developed a simple on-site cataloging tool to remotely gather the collection information. All the information from this cataloging tool feeds into a central database where the data is evaluated to discover customer needs and trends on a global level.

Knowing what records are missing is almost as important as knowing what records are available. It can be very time consuming searching for records long thought to be lost or destroyed. Frequently, alternative sources are needed to replace records no longer available. There are times when critical records are accidentally discovered without there being any record of their existence.

***Figure 1:*** *Rediscovered 200 year old probate records in the basement of a local courthouse*

*Step 2: Create and finalize field projects with record custodians.* All projects start with a signed contract. FamilySearch will be compliant with local laws and customs with regard to

acquiring, storing and publishing records. We use a CRM (Customer Relationship Management) tool to help maintain long standing relationships with archives and donors, some going back almost a hundred years. Content, file formats, publishing rights, image processing and treatments are all specified in the contract and managed in our CRM and process management tools.

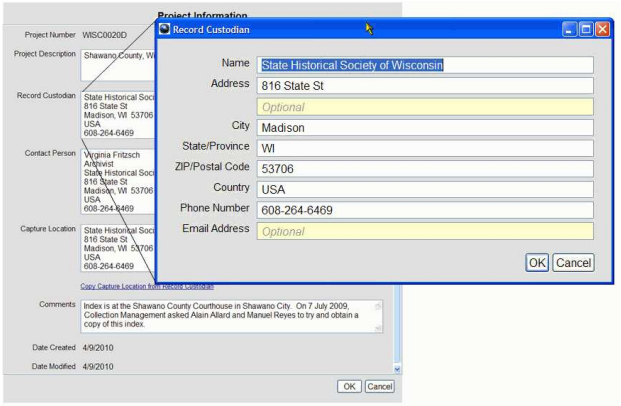


***Figure 2:*** *Detailed archive and contact information*

*Step 3: Field capture.* Publishing collections to the web requires high quality digital images and descriptive metadata that is accurate, consistent and complete. To ensure gathered collections are publishable, FamilySearch developed a tool called dCam-X (digital Camera-X). Through the use of many automated tools, dCam-X enables operators to consistently and effectively capture all the images defined by the contract and produce publishable collections. dCam-X will be discussed in greater detail later in this paper.

*Step 4: Audit and evaluation.* All published collections have been audited by a highly skilled quality assurance team. Images are measured to an "Image Specifications Document" that defines standards such as tonal range, spatial resolution, focus, image blurs, file naming and many other criteria. Folders of images that do not meet our standards are flagged as needing to be reworked

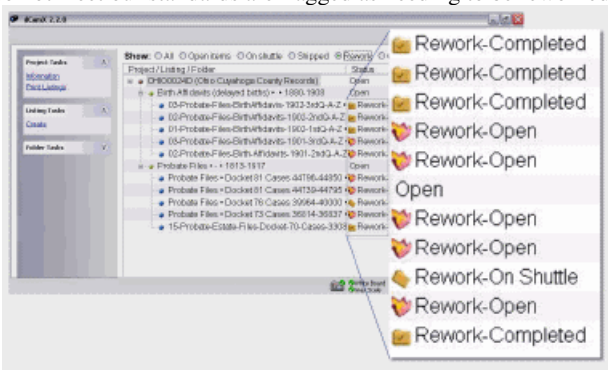


***Figure 3:*** *Management console in dCam-X showing rework for folders that did not pass audit*

and then sent back to the field operator to be recaptured. To make sure content is accurate, complete and compliant with generally accepted cataloging practices, all the metadata is evaluated and enhanced by experienced and certified catalogers.

*Step 5: Publish, preserve and donate:* Along with publishing collections to the web, FamilySearch stores the images in long term preservation along with providing copies of the original images to all archives from which the collections originate. Preservation involves storing multiple copies of each collection at different geographic locations and on different forms of digital media such as tape, DVD and spinning hard drives. Donors are given copies of the original images to provide not only another redundant copy but to also allow the donors to preserve and publish the records as they want. Local patron needs frequently require that archives with the technical means to go ahead and publish the records. Publishing the records by FamilySearch is always in compliance with the terms of the contract. There are instances where FamilySearch will only make indexed data available with links directly to the images manage and served up directly by the archive.

## Field Capture: dCam-X

dCam-X is a portable tool created by FamilySearch to capture collections from wherever they exist throughout the world. Field capture is the most critical step in the publishing process because this is where all the imaging and metadata standards are applied. dCam-X is a semi-automated tool that is designed to allow operators to focus on creating high quality images and not get caught up in the day-to -day management of the files on the system or the overall project. The following is a review of primary features and processes in dCam-X.

*Project Management:* Managing projects from dCam-X allows operators to make sure that nothing is forgotten or overlooked. All project information is electronically managed by dCam-X. Project listings (the list of records defined by the project) are reviewed by experienced catalogers and information specialists before the project is sent out the field operator it is reviewed for

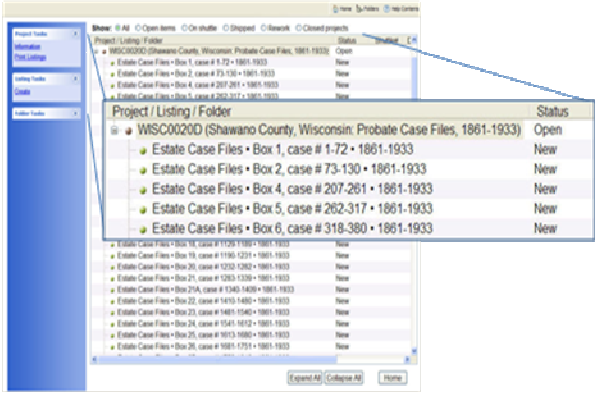


***Figure 4:*** *Management console showing the list of records to be captured as defined by the contract*

accuracy and format. Once a project is received, the field operator

is responsible for going through the listing to make sure that all the records at the archive are available and determine if any minor changes are needed. In the case where records were overlooked during the initial investigation phase, the record item can be added to the listing. All changes to the listing are reviewed to make sure they are compliant with all the terms of the contract.

*Validate record metadata*: the objective of this step is to gather enough information about the records so that a patron is able to search for the collection and then easily browse to the record that contains the information they are looking for. As an operator is preparing to capture a series of records, they will

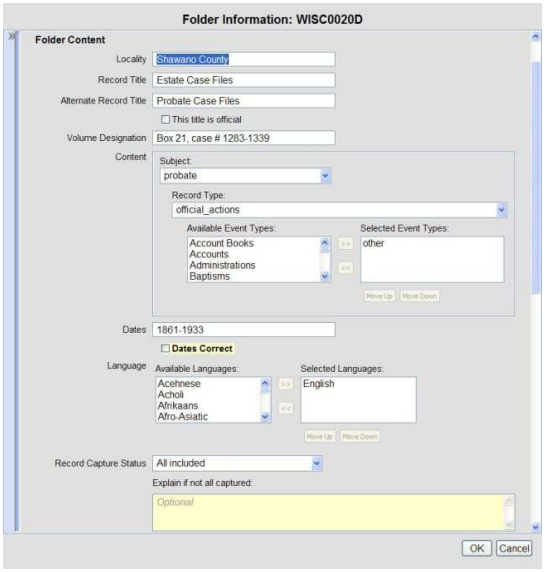


***Figure 5:*** *Some of the fields that need to be evaluated by the camera operator*

review all the metadata associated with the records to make sure that all the information is accurate and complete. Because the operator has the documents in front of them, they are best able to make sure nothing is missing, misspelled or improperly described. Any changes or enhancements to the metadata are passed along with the images for review.

*Camera calibration*: At the beginning of each day and several times throughout the day, dCam-X will walk the camera operators

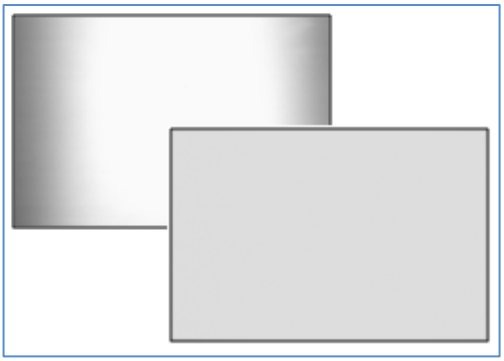

***Figure 6****: The before and after affects of normalizing the array with the whiteboard*

through the calibration process. An experienced operator will take about two minutes to calibrate the system. The steps are as follows:

1. *White-boarding*: Normalizing the CCD-Array to ensure an even expose across the image. Normalization is done by taking a picture of a uniformly reflective white board, measuring the variations in lighting and then computing a corrective mask which is applied to all subsequent images. The value of this to the customer is an image that is evenly readable.

2. *Grayscale*: Grayscale calibration is used to ensure that we are capturing the complete tonal range of the artifact. Operators need to be able to distinguish all the variations of gray ranging from the brightest whites in the document to the darkest backs. Calibration is done by capturing an image of a Kodak grayscale target and then pointing out to dCam-X the A, M and 19 squares on the target. dCam-X will then automatically adjust the system to ensure the proper tonal range of the system. The benefit of this step to the customer is that information on the digital image is not lost. In cases

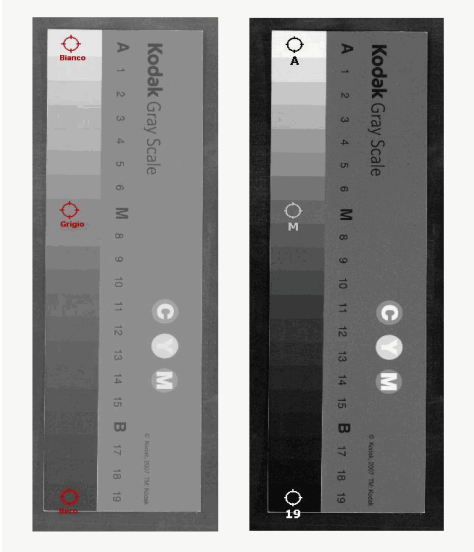


***Figure 7:*** *The before and after affects of applying the grayscale calibration*

where the dark ink on a record is hard to see on the dark background of the aged parchment. Without tonal calibration, the writing won't be seen in the digital image.

3. *Focus*: dCam-X provides a semi-automated method for manually focusing the digital camera. This method uses a target and a focusing algorithm to help the camera operator to accurately focus the camera. As the camera is brought into focus, two lines on the chart, line up and a

checkmark appears when the camera is accurately focused.

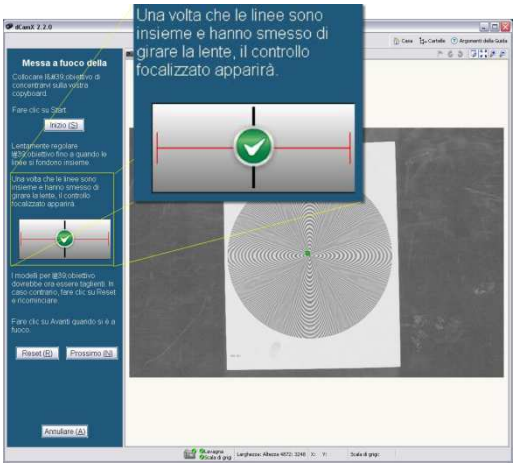


***Figure 8:*** *Focusing tool that will help let the operator know that image is in focus*

4. *LSI*: LSI (line segment indicator) is a tool in dCam-X that helps the operator guarantee the document is being captured at the right resolution. LSI uses the PPLS (pixels-per-line-segment) standard for determining the correct resolution. The PPLS standard requires that the widths of thinnest lines in a document are defined in the digital image by at least three or more pixels. If there are fewer than three pixels, the operator lowers the camera until that standard is met. On the other hand, over sampling will capture all the information on the record but will create unnecessarily large images and waste storage space.

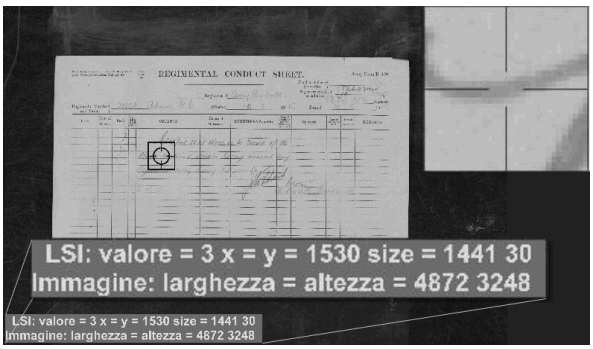


***Figure 9:*** *This image shows that the selected line is at least 3 pixels wide at the point on the line shown in the upper right hand corner*

*Taking Pictures:* dCam-X is designed to digitize original documents in bound or loose formats. Many operators are volunteers in their 60's and 70's, requiring dCam-X to accommodate their special needs. dCam-X will allow a non-technical person with aging eyesight and less than perfectly steady hands to easily create publishable images. Some operators are able to capture documents at a rate of a thousand images per hour or more with few to no mistakes. Where mistakes are made, corrections are fast and easy. Inserts, deletions, retakes and image re-ordering are all done with ease. Corrections to the images or the metadata can be made right up to the point that the images are sent off for audit and processing. If there are needs to make changes after the shuttle has been shipped, we are able to have the field operator send in the corrected folder of images that will replace the defective folder.

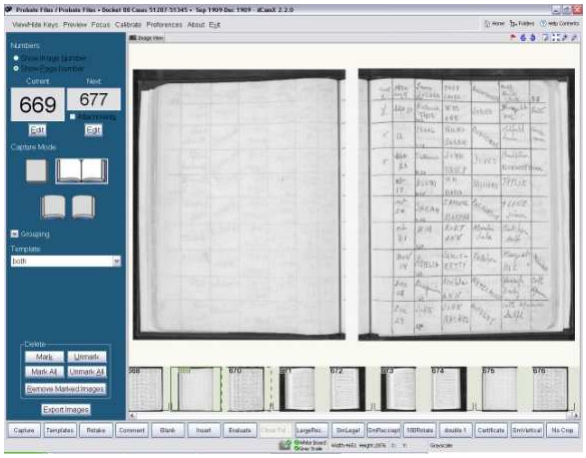


***Figure 10:*** *This figures shows the main working page for dCam-X*

*Evaluate*: operators are required to evaluate 100% of the images before they are sent in for processing. The software will automatically run through all the images, one at a time, allowing the operator to catch any mistakes not corrected during the original capture. Again, the operators are expected to produce publishable quality image and metadata. This step is the last chance an operator has to catch everything.

*Shipping and Data management*: FamilySearch will not only produce consistently high quality images and metadata, but it will also guarantee image integrity throughout all image processing and publishing processes. Images sent back to the donor and those stored in preservation are exact duplicates of the original images captured by the camera. At the time an image is captured and before it is saved to the hard drive, dCam-X will create and store an MD5 hash of both the bit level image and the saved formatted image. Any time the formatted image is shipped for processing, a check of the MD5 is made to ensure that no alterations have been made to the image, either by accident or design. If any changes have been made, dCam-X will not ship the image or any other image in the immediate folder until the operator retakes the altered image. The bit level MD5 hash ensures that the integrity of the images is intact even if the image is transformed between loss-less image formats such as PNG, jpeg2000 or TIFF.

## Conclusion

FamilySearch has implemented processes, tools and standards to ensure that the world's collections are gathered and published as accurately and rapidly as possible. The rate of publishing a million images per week is but a starting point. By expanding our tools and improving our process to meet the global needs of archives and record donors, we hope to eventually be able to grow and capture vital records at the rate they are actually being created.

## FamilySearch Standards Involvement

FamilySearch is a pioneer in national and international standards development for both microfilm and digital imaging technology. FamilySearch representatives served on and chaired the standards development committees for the Association for Information and Image Management/American National Standards Institute (AIIM/ANSI) and the International Organization for Standards (ISO). Today FamilySearch representatives are involved in the following standards committees:

- AIIM/ANSI standards program
    - Chair, C24 Electronic imaging
    - Member of other committees
    - Member of Standard's Boards
- ISO Standards program
    - Member of TC 171, Document Management Applications

FamilySearch also develops internal standards such as the FamilySearch Digital Imaging Specification. This standard is used to establish and enforce specifications for image format, quality and delivery. Internal standards and specifications developed by FamilySearch have been and will continue to be proposed for consideration and adoption by national and international standards organizations for use as industry standards.

## About FamilySearch

FamilySearch is the largest genealogy organization in the world. Millions of people use FamilySearch records, resources and services to learn more about their family history. To help in this great pursuit, FamilySearch has been actively gathering, preserving and sharing genealogical records worldwide for over 100 years. FamilySearch is a nonprofit organization sponsored by the Church or Jesus Christ of Latter-day Saints. Patrons may access FamilySearch services and resources free online at familysearch.org or through over 4,500 family history centers in 70 countries, including the main Family History Library in Salt Lake City, Utah

## Author Biography

*Paul D. Abbott received his BS is Mechanical Engineering from Brigham Young University and his Masters in Engineering Management (MEM) from Brigham Young University. He is currently responsible for defining and developing the next generation of digital tools and processes needed to rapidly gather and publish collections at FamilySearch.*